CATEGORICAL CVA BIPLOTS

David Timothy Rodwell

Assignment presented in partial furfillment
of the requirement for the degree
MCom (Financial Risk Management)
at the University of Stellenbosch

Supervisor: Dr. C.J. van der Merwe

Degree of confidentiality: A

December 2020

ACKNOWLEDGEMENTS

I want to use this opportunity to thank all those who contributed to the success of this assignment. However, I wish to give a special mention to the following individuals:

- Firstly, to my parents and sister for their never-ending support and love throughout my academic career.
- To my supervisor, Dr. van der Merwe, you have shown me what the process of learning ought to be. Thank you for listening and guiding me through this assignment.
- To Prof. Lubbe for her time and effort, she set out to improve the readability and presentation of this assignment.
- My close friends and roommates who made the journey a memorable one.

SUMMARY

In the modern era a great amount of emphasis is placed on data visualisation, especially in cases where a large amount of data is present. Usually, in these instances, the data is of a high-dimensional nature which cannot be visualised using conventional means. Fortunately, there has been a recent surge in using biplots to visualise multivariate data, where biplots can be described as a generalisation of a scatterplot. Moreover, these biplots use dimension reduction techniques to construct a two-dimensional representation of the data with non-orthogonal axes. However, at present, an effective biplot construction technique which adequately separates classes, in cases where categorical data is present does not exist.

Hence, this research builds upon an existing biplot construction technique by using elements from Canonical Variate Analysis (CVA) and non-linear Principal Component Analysis (PCA) to develop a technique that can perform class separation in cases where numerical and categorical data is present. This novel biplot construction methodology forms the crux of this research assignment. Subsequently, the feasibility of this method was explored by considering the well-known Iris data set where two variables are binned to form categorical variables. It is shown that this novel method improves upon existing biplot construction in terms of classification accuracy and class separation. However, it is noted this method can be extended by incorporating CVA in the iterative algorithm which solves the optimal categorical level scores.

A web-based Shiny application was built as supplement to this paper, and can be found at https://davidrodwell.shinyapps.io/CategoricalCVABiplotApp/. Here the user can interact with the data sets, proposed methodology, and functionalities presented in this research.

OPSOMMING

In die moderne era word daar baie klem gelê op die visualisering van data, veral in waar groot datastelle betrokke is. In hierdie gevalle is die data gewoonlik hoë-dimensioneel van aard, wat veroorsaak dat dit nie deur konvensionele maniere visueel voorgestel kan word nie. Onlangse verwikkelinge het gelei tot 'n toename in die gebruik van bi-stippings om multi-veranderlike data voor te stel, waar bi-stippings as 'n veralgemening van 'n spreidingsdiagram beskryf kan word. Hierdie bi-stippings gebruik dimensie verminderingstegnieke om 'n twee-dimensionele voorstelling van die data op 'n nie-ortogonale assestelsel te konstrueer. Huidiglik bestaan daar nie 'n effektiewe bi-stipping konstruksietegniek wat klasse kan verdeel wanneer kategoriese data teenwoordig is nie.

Hierdie navorsing bou op 'n bestaande bi-stipping konstruksietegniek wat elemente van Kanoniese Veranderlike Analise (KVA) en nie-lineêre Hoof Komponent Analise (HKA) gebruik om 'n tegniek te ontwikkel wat klasse kan verdeel in gevalle waar numeriese sowel as kategoriese data teenwoordig is. Hierdie nuwe bi-stipping konstruksie metodologie vorm die kruks van hierdie navorsingstaak. Die lewensvatbaarheid van hierdie metode was ook ondersoek deur die welbekende Iris datastel te oorweeg, waar twee veranderlikes ingedeel word om kategoriese veranderlikes te word. Dit is gewys dat hierdie nuwe metode die bestaande biplot konstruksietegnieke verbeter in terme van klassifikasie akkuraatheid en klas verdeling. Daar was wel opgemerk dat hierdie metode uitgebrei kan word deur KVA in die iteratiewe algoritme te inkorporeer, wat die optimale kategoriese vlak tellings oplos.

'n Web-gebaseerde Shiny toepassing was gebou as supplimentêr tot hierdie artikel, en kan gevind word by https://davidrodwell.shinyapps.io/CategoricalCVABiplotApp/. Hier kan die gebruiker self interaksie hê met die datastelle, voorgestelde metadologie, en funksionaliteite wat voorgelê is in hierdie navorsing.

SAQA OUTCOMES

For this masters assignment an article based approach was followed. This entailed producing a working paper (provided in this assignment document) which was given a provisional mark (on submitting to a journal) and corrections by the examiner. These corrections were incorporated and the paper was submitted to a journal that was agreed upon by the student and supervisor. To ensure compliance of the outcomes of the assignment, a list of all the required South African Qualification Authority (SAQA) outcomes[1] are listed below and how they were achieved in this assignment.

Outcome	Fulfilment
Scope of knowledge	A novel method of visualising multivariate data using categorical CVA biplots, is developed.
Knowledge literacy	Literature and code from existing methods were extensively examined to develop the new biplot construction method in similar coding environments.
Method and procedure	A proof of concept was developed using a well-known data set. Also, all underlying theory and processes are presented in detail.
Problem-solving	A wide range of tools were used in developing this biplot construction method namely, multivariate dimensional scaling, coding in `R`, linear algebra and implementing advanced visualisation techniques which allow for easier interpretation of categorical data. Furthermore, the consequences of this biplot construction method pertaining to more effective classification are mentioned.
Ethics and professional practice	Careful consideration was given in obtaining and using the data. Professionalism was upheld to the highest standard at all times. Furthermore, all required ethical clearances were obtained where necessary.
Accessing, processing and managing information	An extensive literature review and methodology section are given. The literature review provides an overview of current methods of multivariate visualisation in the context of biplots, with the methodology section demonstrating how the new method will expand upon existing research.
Producing and communicating information	The learner took part in an Elseriver course where items such as the communication of skills and ideas were presented. The target audience was always taken into account during the writing of both pieces.
Context and systems	Several areas of existing code had to be amended to effectively produce biplots using the new methodology.
Management of learning	Several consulting sessions with the learner's supervisor aided in promoting self learning strategies. These sessions allowed the learner to gain independence with regards to his learning and enabled him to enhance his skills to work in a professional environment.
Accountability	Both papers were written independently by the learner incorporating edits suggested under the guidance of his supervisor.

[1] See: https://www.saqa.org.za/docs/misc/2012/level_descriptors.pdf

PLAGIARISM DECLARATION

1. Plagiarism is the use of ideas, material and other intellectual property of another's work and to present it as my own.
2. I agree that plagiarism is a punishable offence because it constitutes theft.
3. Accordingly, all quotations and contributions from any source whatsoever (including the internet) have been cited fully. I understand that the reproduction of text without quotation marks (even when the source is cited) is plagiarism.
4. I also understand that direct translations are plagiarism.
5. I declare that the work contained in this assignment, except otherwise stated, is my original work and that I have not previously (in its entirety or in part) submitted it for grading in this assignment or another assignment.

Student number	Signature
D.T. Rodwell	17 October 2020
Initials and surname	Date

CATEGORICAL CVA BIPLOTS

A shorted version of this paper was submitted to an international statistical journal.

Abstract

In a world where data is becoming one of the sought after assets, techniques to visualise and understand large amounts of data is paramount. In most settings, this data is usually of a high-dimensional nature which further stresses the need for effective visualisation techniques. Hence, this paper expands upon a multivariate visualisation technique called biplots in cases where categorical variables are present. In particular, a new biplot construction methodology, named CVA(H_r), which incorporates concepts from both non-linear principal component analysis and canonical variate analysis, is developed. This technique is then showcased using the Iris data set where two variables are binned to form categorical variables. It is shown that this novel method improves upon existing biplot construction in terms of classification accuracy and class separation.

Keywords: Biplots, CVA, Categorical Data

TABLE OF CONTENTS

1. Introduction

In the wake of the industrial revolution, the introduction of computers drastically transformed the way in which data is processed and visualised. A large emphasis has subsequently been placed on highly flexible classification methods to effectively analyse these vast amounts of data. In using these methods, however, the interpretability of the resulting coefficients is usually either infeasible or difficult. This well-known trade-off between flexibility and interpretability is visually described in James, Witten, Hastie and Tibshirani (2013). Most classification techniques do not provide output in a form that is easily comprehensible. This issue is further compounded in cases where the underlying data is highly-dimensional. Van der Merwe (2020) states that in cases where the visualisation of a classification technique is possible, the resulting classification will be better understood than compared to complex "black-box" methods.

This paper expands upon an existing multivariate visualisation technique known as biplots, where biplots can be best understood as a multivariate generalisation of a simplistic scatterplot (Greenacre, 2010). In particular, this paper assesses the feasibility of using non-linear Principal Component Analysis (PCA) in conjunction with Canonical Variate Analysis (CVA) when categorical variables are present. The effectiveness of this biplot construction method is demonstrated through the use of a transformed version of the well-known Iris data set where two variables were transformed into categorical variables using a binning procedure.

The remainder of the paper is as follows: the construction of PCA, CVA, and non-linear PCA biplots are discussed in section 2; thereafter the new technique, referred to as CVA(H_r), will be presented along with the IRIS data set, which will be used to construct the biplots. Thereafter, the various sets of biplots namely, catPCA and CVA(H_r) will be given together with accuracy metrics in section 4 to show that the new CVA(H_r) technique improves on the existing non-linear PCA biplot; the paper is concluded in section 5 with a summary and discussion of areas for further research.

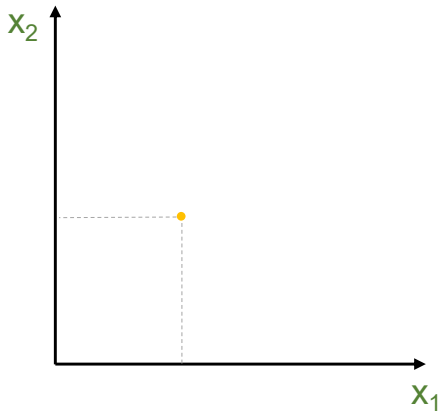
2. Literature Review

In this section some necessary mathematical background on the construction of the two most popular biplots, namely PCA and CVA, are provided. The PCA biplot has further been extended to a non-linear PCA case, which can accommodate both nominal and ordinal categorical variables. The necessary literature overview and mathematical background for this extension is also provided. These three biplots forms the basis of the methodology introduced in section 3.

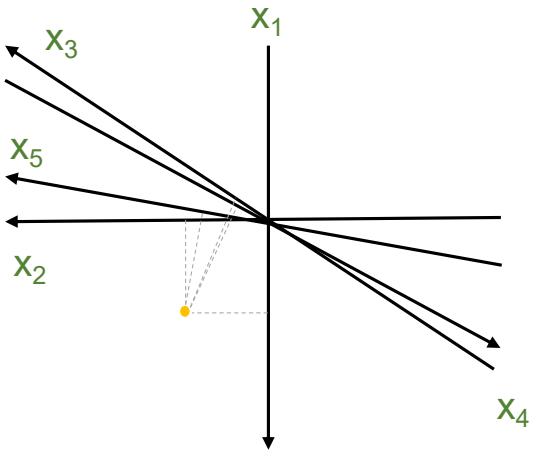
2.1. Biplots

In the traditional setting of two variables, a scatterplot can be utilised to visually represent data using two orthogonal axes. As mentioned in Greenacre (2010), biplots can be seen as an extension of the scatterplot to accommodate p variables by introducing p axes which can

assume any orientation. In the same manner that the values of the two variables associated to a particular point can be read on a scatterplot by perpendicularly projecting points onto the X_1 and X_2 axis, the values of the p variables can be read in a biplot by perpendicular projecting the points onto the p axes, as demonstrated in figure 1. It should be noted that these values are only an approximation since the dimension of the biplot is lower than the true dimensionality of the data. Furthermore, in a biplot, the correlation between two variables can be inferred by analysing the direction of the axes. Two variables will have a strong positive correlation if their axes point in the same direction (X_2 and X_5 in figure 1(b)), whereas a strong negative correlation will be apparent if the two axes point in opposite directions (X_3 and X_4 in figure 1(b)). If two axes are perpendicular to each other, then one can infer that the variables show little or weak correlation (X_1 and X_2 in figure 1(b)).



(a) A scatterplot representing the relationship between X_1 and X_2.



(b) A biplot representing the relationship between five variables with non-perpendicular axes.

Figure 1: A comparison between a traditional scatterplot and a biplot adapted from Greenacre (2010).

The remainder of this section is dedicated to presenting the various techniques used to construct three variations of biplots, namely PCA, CVA, and non-linear PCA types.

2.2. PCA

When visualising high dimensional data in two or three-dimensions an effective dimension reduction technique is needed. Fortunately, many such dimension reduction techniques exist in literature and one of the most well-known techniques is that of PCA. One of the key benefits of using PCA in this setting is that it is easy to understand and explain to non-practitioners. The aim of PCA is to obtain a set of uncorrelated linear combinations of measured variables that can best summarise the total sample variable from a larger set of measured variables (Hotelling, 1933). A simple and coherent summary of PCA can be found in Jolliffe and Cadima (2016) and is summarised as follows.

Let $\mathbf{X} : n \times p$ be a centred matrix such that $\mathbf{1}'\mathbf{X} = \mathbf{0}'$. Then PCA involves finding any linear combination such that $\text{Var}(\mathbf{Xm}) = \mathbf{m}'\mathbf{Sm}$ is maximised, where $\mathbf{m}$ is a vector of constants

$m_1, m_2, \ldots, m_p$ and $\mathbf{S}$ is the sample covariance matrix. In order to maximise the above problem with a well-defined solution, an additional restriction is imposed, namely that the vector of constants are unit-norm vectors ($\mathbf{m}'\mathbf{m} = \mathbf{1}$). Hence, the problem is equivalent to maximising the following expression,

$$\mathbf{m}'\mathbf{S}\mathbf{m} - \lambda(\mathbf{m}'\mathbf{m} - 1) \tag{1}$$

where λ is a La Grange multiplier. After differentiating (1) with respect to the vector $\mathbf{m}$ and setting equal to the null vector, it reduces to the following eigenvector equation

$$\mathbf{S}\mathbf{m} = \lambda\mathbf{m}. \tag{2}$$

Thus, the solution to (2) is $\mathbf{m}$, the eigenvectors of $\mathbf{S}$. Since $\mathbf{S}$ is a real symmetric matrix, it follows that $\mathbf{S}$ has exactly p real eigenvalues. In order to satisfy the additional constraint, the eigenvectors can be normalised to form an orthonormal set. The columns of $\mathbf{X}\mathbf{m}$, with $\mathbf{m}$ now denoting the set of orthonormal eigenvectors of $\mathbf{S}$, are commonly referred to as principal components. Hence, it follows that the full set of eigenvectors of $\mathbf{S}$ are the solutions to the problem presented in (1). In the following subsection, it will be shown that performing singular value decomposition (SVD) on the centred matrix is equivalent to PCA.

2.2.1. PCA and its relation to SVD

By definition, any arbitrary matrix $\mathbf{X} : n \times p$ of rank r, where $r \leq \min(n, p)$, can be written as,

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}', \tag{3}$$

where $\mathbf{U} : n \times r$, $\mathbf{V} : p \times r$ are matrices with orthonormal columns such that $\mathbf{U}'\mathbf{U} = \mathbf{I} = \mathbf{V}'\mathbf{V}$ and $\mathbf{D} : r \times r$ a diagonal matrix. The elements of $\mathbf{D}$ are equal to the non-negative square root eigenvalues, otherwise known as the singular values, of $\mathbf{X}$ in decreasing order.

Furthermore, suppose $\mathbf{X}$ is a centered matrix so that $\mathbf{X}'\mathbf{X} = (n-1)\mathbf{S}$, then by (3) it is clear that,

$$\mathbf{S} = \frac{1}{n-1}\mathbf{X}'\mathbf{X} = \frac{1}{n-1}(\mathbf{U}\mathbf{D}\mathbf{V}')'(\mathbf{U}\mathbf{D}\mathbf{V}') = \frac{1}{n-1}\mathbf{V}\mathbf{D}\mathbf{U}'\mathbf{U}\mathbf{D}\mathbf{V}' = \mathbf{V}\frac{\mathbf{D}^2}{n-1}\mathbf{V}'. \tag{4}$$

Comparing this with the eigenvector problem in (2), the eigenvalues of $\mathbf{S}$ are on the diagonal of the matrix $\frac{\mathbf{D}^2}{n-1}$, and the columns of $\mathbf{V}$ are the corresponding eigenvectors. Therefore, the principal components in the case of SVD are given by $\mathbf{X}\mathbf{V}$. Thus, the results given above show that SVD on the centred matrix is equivalent to PCA. In the next subsection it is demonstrated how SVD is used to construct a PCA biplot.

2.2.2. PCA biplot construction

The methodology on how to construct arguably the most traditional biplot, namely the PCA biplot, is presented in Gower, Lubbe and Le Roux (2011), and can be summarised as follows. Let $\mathbf{X} : n \times p$ be a centred matrix of rank r. Then $\mathbf{X}$ can be written in terms of its SVD as $\mathbf{X} = \mathbf{UDV}$, with $\mathbf{U} : n \times r$, $\mathbf{D} : r \times r$ and $\mathbf{V} : r \times p$, such that the columns of $\mathbf{U}$ and $\mathbf{V}$ are orthonormal. Thereafter, the best rth dimensional representation of the original data set $\mathbf{X}$ is obtained. Usually r is chosen to be two, which then requires a set of coordinates. The coordinates, say $\{Z_1, Z_2\}$, of the best-fitting two-dimensional approximation of the centred matrix $\mathbf{X}$ is given by $\mathbf{Z} = \mathbf{X}\mathbf{V}_{[\mathbf{2}]}^{'}$, where $\mathbf{V}_{[\mathbf{2}]} : p \times 2$ are any two rows of $\mathbf{V}^{'}$ (usually the first two as they relate to the largest eigenvalues). Moreover, the two rows of $\mathbf{V}_{[\mathbf{2}]}^{'}$ are used in determining the direction of the axes and providing the coordinates of unit markers. However, as noted in Gower and Hand (1996), these markers are not scaled to the standardised variable for which additional multivariate scaling techniques are used to extend the axes to cover the full plotting space and to have the accompanying tick marks.

2.3. CVA

The journey to CVA began with the more familiar linear discriminant analysis (LDA) proposed by Fisher (1936). This involved finding a linear combination of p-measured variables which best discriminates between two groups. The manner in which LDA best discriminates between groups is to find a vector $\mathbf{m}$ such that the ratio of the total variance ($\mathbf{\Sigma}$) to the within-group variance ($\mathbf{\Sigma_W}$) is maximised i.e.

$$\max_{\mathbf{m}} \frac{\mathbf{m}^{'}\mathbf{\Sigma}\mathbf{m}}{\mathbf{m}^{'}\mathbf{\Sigma_W}\mathbf{m}}. \tag{5}$$

A linear combination $\mathbf{m}^{'}\mathbf{x}$ which maximises (5) is known as the first linear discriminant function or the first canonical variable and will form the complete solution for the two-group case.

Thereafter, Rao (1948, 1952), proposed an extension of Fisher's LDA to the multi-group case. The aim in this multi-group setting is to obtain a set of uncorrelated linear combinations of p-measured variables that best discriminate between J groups. Now suppose that $\mathbf{X} : n \times p$ is the matrix of individual observations which are assumed to be centered such that $\mathbf{1}^{'}\mathbf{X} = \mathbf{0}^{'}$ so that $\mathbf{X}^{'}\mathbf{X}$ is the total sums of squares and cross-products (SSP) matrix. Further, suppose that $\mathbf{X}^{'}\mathbf{X}$ can be partitioned into a between-groups SSP matrix, $\mathbf{B}$, and a within-group SSP matrix $\mathbf{W}$ such that $\mathbf{X}^{'}\mathbf{X} = \mathbf{B} + \mathbf{W}$. Then the linear combination which optimally discriminates between J groups in the case of n samples is defined by the vector $\mathbf{m}$ which maximises the sample variance ratio,

$$\frac{\mathbf{m}^{'}\widehat{\mathbf{\Sigma}}\mathbf{m}}{\mathbf{m}^{'}\widehat{\mathbf{\Sigma}}_{\mathbf{W}}\mathbf{m}}. \tag{6}$$

It can further be shown that since $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{W}}$ are proportional to $\mathbf{X}'\mathbf{X}$ and $\mathbf{W}$, that the vector $\mathbf{m}$ also maximises the following ratio,

$$\frac{\mathbf{m}'\mathbf{X}'\mathbf{X}\mathbf{m}}{\mathbf{m}'\mathbf{W}\mathbf{m}}. \tag{7}$$

Moreover, taking into account $\mathbf{X}'\mathbf{X} = \mathbf{B} + \mathbf{W}$, a vector that maximises the ratio,

$$\frac{\mathbf{m}'\mathbf{B}\mathbf{m}}{\mathbf{m}'\mathbf{W}\mathbf{m}}, \tag{8}$$

also maximises the ratio given in (7). For the rest of this section only the ratio given in (8) will be considered. Since the vector $\mathbf{m}$ is only uniquely defined up to a scalar multiple, the search space is subsequently restricted by imposing a constraint, namely that $\mathbf{m}'\mathbf{W}\mathbf{m} = \mathbf{1}$. Thus the vector $\mathbf{m}$ which maximises (8) while satisfying the constraint above as given in Gower and Hand (1996), is a non-zero solution of the following equation:

$$\frac{d}{d\mathbf{m}}\Big(\mathbf{m}'\mathbf{B}\mathbf{m} - \lambda(\mathbf{m}'\mathbf{W}\mathbf{m} - \mathbf{1})\Big) = 0 \tag{9}$$

$$\Rightarrow \mathbf{B}\mathbf{m} = \lambda\mathbf{W}\mathbf{m}. \tag{10}$$

Hence, the vector $\mathbf{m}$ which maximises the ratio in (8) is the eigenvector of the two-sided eigenvalue problem. It is clear that the p eigenvectors and eigenvalues can be represented simultaneously by,

$$\mathbf{B}\mathbf{M} = \mathbf{W}\mathbf{M}\boldsymbol{\Lambda}, \tag{11}$$

where $\boldsymbol{\Lambda} : p \times p$ is a diagonal matrix with the ith diagonal element equal to the ith eigenvalue obtained in the two-sided eigenvector problem satisfying the constraint $\mathbf{M}'\mathbf{W}\mathbf{M} = \mathbf{I}$. This constraint can be viewed as the initial constraint extended over the p measurable variables. Furthermore, Rao (1952) notes that the ith eigenvalue (λ_i) in the diagonal of $\boldsymbol{\Lambda}$ quantifies how effective the linear combination $\mathbf{m}_i'\mathbf{x}$ separates the groups, where $\mathbf{m}_i$ is the ith row of $\mathbf{M}$. Subsequently, the coefficient vector $\mathbf{m}_i : i = 1, \ldots, p$ is defined as the ith sample canonical variable. However, since the last $p - K$, where $K = \min(J - 1, p)$, coefficient vectors are not uniquely defined, only the first K coefficient vectors will be considered and be referred to as canonical variables. Hence, the K rows spanned by $\mathbf{X}\mathbf{M}$ will occupy at most $K = \min(J-1, p)$ dimensions and may be approximated in fewer dimensions by PCA.

The above gives a clear indication that canonical variables for a sample can be obtained by applying a non-singular linear transformation $\mathbf{M}$ on $\mathbf{x}$ such that the rows spanned by,

$$\mathbf{y}' = \mathbf{x}'\mathbf{M}, \tag{12}$$

are canonical variables. Moreover, Gower and Hand (1996) states that any pair of coefficient vectors, $\mathbf{m}_i$, $\mathbf{m}_j$ with $j > i$ and $i, j \in [1, K]$, are orthogonal so that the group separation is performed in opposite directions resulting in canonical variables being uncorrelated with

each other. Since canonical variables can be perfectly represented in $K \leq p$ dimensions, in addition to the canonical variables being chosen to maximise the group separation while being uncorrelated with each other, the multi-group case of LDA can be considered equivalent to CVA. It must be noted that this approach of CVA is made possible through a one-step method by considering the solution to the two-sided eigenvalue problem. The next method discussed will consider a two-step approach as proposed by Gower et al. (2011).

The two-step approach provided by Gower et al. (2011) first considers the transformation of the original variables into the canonical space and thereafter, approximating the canonical sample points by PCA. In the multi-group LDA setting, these two steps were combined by solving the two-sided eigenvalue problem.

A transformation into the canonical space as noted by Gower et al. (2011) is governed by a nonsingular transformation matrix $\mathbf{L} : p \times p$ such that $\mathbf{L}\mathbf{L}^{'} = \mathbf{W}^{-1}$, which can be obtained as the solution to the following eigenvalue equation,

$$\mathbf{W}\mathbf{L} = \mathbf{L}\mathbf{\Lambda}. \tag{13}$$

Thereafter, the eigenvectors of $\mathbf{L}$ are scaled so $\mathbf{L}^{'}\mathbf{W}\mathbf{L} = \mathbf{I}$. This scaling ensures that all eigenvectors are orthogonal to each other. As a result, the transformed values $\mathbf{y}^{'} = \mathbf{x}^{'}\mathbf{L}$ of the p-measured variables reside in the canonical space and are referred to as canonical variables which are uncorrelated with each other.

In the second step, PCA is performed on the canonical means $\bar{\mathbf{X}}\mathbf{L}$ through the use of SVD. In particular, the PCA approximation of the canonical means $\hat{\bar{\mathbf{X}}}\mathbf{L}$ is $(\bar{\mathbf{X}}\mathbf{L})\mathbf{V}\mathbf{J}\mathbf{V}^{'}$ where $\mathbf{J} : p \times p$ is the p-dimensional unit matrix and $\mathbf{L}\mathbf{V} = \mathbf{M}$ as in the case of the multi-group LDA. This application of PCA can be geometrically interpreted as fitting a plane, or hyperplane in the case of $K > 3$ dimensions, of best fit through the canonical means. The benefit of using SVD in this instance is that it inherently maximises the variance ratio given in (8) as a natural consequence of the least-squares property in the SVD.

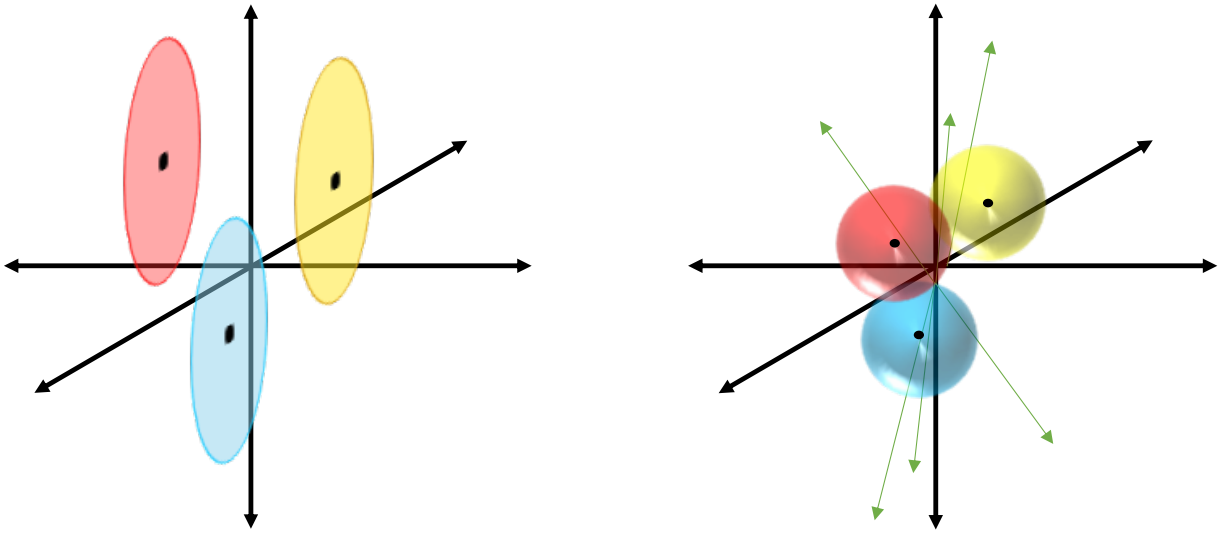
It is clear that the two-step approach provides a method of generating canonical variables in a reduced number of dimensions that optimally separates groups. Furthermore, these canonical variables is obtained in such a manner so that each of the canonical variables are uncorrelated with each other. Hence, the two-step approach can be considered as another case of CVA.

2.3.1. CVA biplot construction

As mentioned in Campbell and Atchley (1981), CVA biplots provides a method of visualisation to represent differences between the means of K groups in a reduced number of dimensions. The application of using CVA biplots on multivariate data has been explored in multiple disciplines, for example Gardner and Le Roux (2005), Walters and Le Roux (2008), Aldrich et al. (2004), Varas et al. (2005) and Alkan et al. (2015)) amongst others.

Gower et al. (2011) describes the manner in which a CVA biplot is constructed using the two-step approach. The first step involves transforming the target matrix $\mathbf{X} : n \times p$ into the

canonical space $\mathbf{Y} = \mathbf{XL}$ such that $\mathbf{L}'\mathbf{WL} = \mathbf{I}$. This transformation is visually explained in figure 2(b), where the green arrows represent the rotation of the old axes. The nonsingular matrix $\mathbf{L}$ scales the eigenvectors so that the Mahalanobis[2] D^2 distance between-class means in the original p-dimensional space are Pythagorean[3] distances in the canonical space. Moreover, the assumption that the independent variables are normally distributed under each group, results in the classification regions assuming the form of a r-dimensional sphere. In the case of three classes having ellipsoidal distribution the transformation to the canonical space can be seen in figure 2(b).



(a) A figure of three classes with ellipsoidal shaped distributions with the class means assumed centred around the origin. The ellipse around the points denote the classification regions (individual sample points not shown).

(b) A figure of three transformed classes transformed to the canonical space through the relationship $\mathbf{XL}$. The ellipsoids around the points denote the classification regions.

Figure 2: A figure adapted from Gower et al. (2011), depicting the transformation of three classes from the original data matrix $\mathbf{X}$, which is assumed to follow a ellipsoidal shaped distribution, into the canonical space through the relationship $\mathbf{Y} = \mathbf{XL}$, where $\mathbf{L}$ is a nonsingular matrix. The green arrows represent the relative position of the new axes to the old axes (black), with the various rotations governed by the matrix multiplication $\mathbf{XL}$. The spherical nature of the classification region is a result of the normality assumption of CVA.

After transforming the original variables into the canonical space, PCA is used to fit the least-squares plane, or hyperplane for more than three classes, to the canonical means. In particular, PCA is used to approximate $\bar{\mathbf{X}}\mathbf{L}$, the canonical means, with $\widehat{\mathbf{Y}} = \widehat{\bar{\mathbf{X}}}\mathbf{L} = \mathbf{UDV}_{[\mathbf{2}]}$. The above is shown in figure 3. Thereafter, the class mean coordinates can be approximated in the two-dimensional space by $\bar{\mathbf{Z}} = \bar{\mathbf{Y}}\mathbf{V}'_{[\mathbf{2}]} = \bar{\mathbf{X}}\mathbf{L}\mathbf{V}'_{[\mathbf{2}]} = \bar{\mathbf{X}}\mathbf{M}'_{[\mathbf{2}]}$ where $\mathbf{M}'_{[\mathbf{2}]} : p \times 2$ is the normalised eigenvector of $\mathbf{LV}'_{[\mathbf{2}]}$ and is used in constructing the axes of the respective

[2]The Mahalnobis D^2 distance between class k and h is defined as $\delta_{kh} = \sqrt{(\bar{\mathbf{x}}_k - \bar{\mathbf{x}}_h)'\mathbf{W}^{-1}(\bar{\mathbf{x}}_k - \bar{\mathbf{x}}_h)}$

[3]In this research proposal the Pythagorean distance is the ordinary Euclidean distance with the distances between class means k and h defined as $d_{kh} = \sqrt{(\bar{\mathbf{x}}_k - \bar{\mathbf{x}}_h)(\bar{\mathbf{x}}_k - \bar{\mathbf{x}}_h)}$

variables. The coordinates for the two-dimensional approximation of the canonical subspace is given by $\mathbf{Z} = \mathbf{XLV}'_{[\mathbf{2}]}$. These points are then classified to its nearest canonical mean according to the Mahalanbobis distance metric. Also, since CVA is equivalent to multi-class LDA, CVA optimally discriminates between the J groups by maximising the between-class to the within-class variance ratio given in (8). This is one of the main advantages if using CVA over PCA when constructing biplots.

Although PCA and CVA biplots allow for an effective visualisation technique in the presence of high-dimensional data, they are restricted to only incorporate numerical variables. One method to introduce categorical variables is through label encoding, where each unique value is assigned a number, however this is not ideal. For this reason non-linear PCA biplots were developed. This biplot construction technique aims to convert a data matrix, which consists of numerical and categorical variables, to an optimal numeric form. These non-linear PCA biplots will be the topic covered in the following subsection.

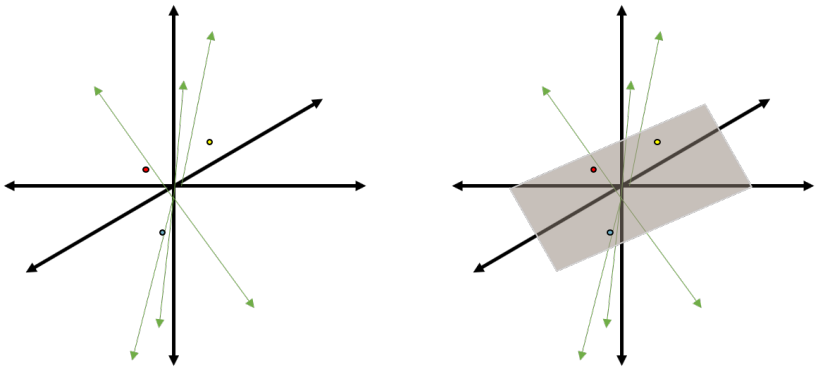


Figure 3: The dots in the left diagram represent the class means in the canonical space. A least-squares plane (grey rectangle) can then be fitted through each of these points, resulting in a 2-dimensional space which forms the basis of the CVA biplot.

2.4. Non-linear PCA

Non-linear PCA biplots, also known as catPCA biplots, is a technique used in cases where categorical variables with different measurement levels (e.g. ordinal and nominal) are present within the original data matrix $\mathbf{X} : n \times p$. As in the case of traditional PCA, the objective of catPCA is to reduce the dimension of the data set $\mathbf{X}$ into a smaller target matrix $\mathbf{Y} : n \times r$ where $r < p$ such that the variables in $\mathbf{Y}$ are uncorrelated but still represent the original data set $\mathbf{X}$.

In order to represent categorical (both nominal and ordinal) variables in a matrix, a pseudo-numeric form is used. This form records the nominal variables such that the kth variable is represented by the matrix $\mathbf{G_k} : n \times L_k$ where L_k denotes the number of category levels. The $\mathbf{G_k}$ matrix is structured in a manner such that the ith row of $\mathbf{G_k}$ is zero except for a single unit in the column relating to the actual category level taken by the ith sample. The

indicator matrix $\mathbf{G}$ for the complete data is obtained by combining the category variables to give,

$$\mathbf{G} = [\mathbf{G_1}, \mathbf{G_2}, \ldots, \mathbf{G_k}, \ldots, \mathbf{G_p}] : n \times L, \tag{14}$$

where $L = L_1 + L_2 + \ldots + L_k + \ldots + L_p$.

Now, suppose the original data matrix $\mathbf{X}$ can be written in the form denoted as,

$$\mathbf{H_{cat}} = [\mathbf{G_1 z_1}, \mathbf{G_2 z_2}, \ldots, \mathbf{G_k z_k}, \ldots, \mathbf{G_p z_p}] \tag{15}$$

where $\mathbf{G}$ is the indicator matrix of complete data and $\mathbf{z_k} : L_k \times 1$ is a numeric column vector, often termed quantifications. As mentioned in Gower et al. (2011), in homogeneity analysis the aim is to replace the category levels by numerical optimal scores such that $\mathbf{X}$ can be replaced by $\mathbf{H_{cat}}$, otherwise stated, finding an optimal numeric representation of $\mathbf{X}$. Hence, the goal of catPCA is to find optimal vectors of $\mathbf{z_k}$ such that it maximises the sum of the first r eigenvalues of the matrix resulting from principal components of $\mathbf{H_{cat}}$. It is worthwhile to note that, in the case of a numeric variable (say $\mathbf{t}$), the initial $\mathbf{z_k}$ vector entries will be the numeric values of the continuous variable so that the kth column of $\mathbf{H_{cat}}$ is equal to $\mathbf{G_k z_k} = \mathbf{It}$, where $\mathbf{I}$ is the usual identity matrix. Assuming that all the $\mathbf{z}$ vectors are scaled, such that centred columns of $\mathbf{H_{cat}}$ are normalised, the objective function for catPCA can be formulated as

$$\min ||\mathbf{H_{cat}} - \mathbf{Y_{cat}}||^2, \tag{16}$$

where $\mathbf{Y_{cat}} = \mathbf{UDV_{[r]}}$ is the SVD of the centred matrix $\mathbf{H_{cat}}$ with $\text{rank}(\mathbf{Y_{cat}}) = r < p$. In this paper only two-dimensional biplots are considered, hence in this step we set $r = 2$. However, when $\mathbf{Y_{cat}}$ is known, the objective function in (16) may be written as follows,

$$\min \sum_{k=1}^{p} ||\mathbf{G}_k \mathbf{z}_k - \mathbf{y}_k||^2 \tag{17}$$

where $\mathbf{y}_k$ is the k-th column of $\mathbf{Y}$. Considering the representation of the objective function in (17), the optimal $\mathbf{z}$ quantifications can be obtained independently by solving,

$$\min ||\mathbf{G}_k \mathbf{z}_k - \mathbf{y}_k||^2. \tag{18}$$

It is noted, as stated in Gower et al. (2011), that $\mathbf{Y}$ can be rearranged in order to ensure that it has zero column sums. Hence, only the constraint on $\mathbf{z}_k$, namely, $\mathbf{z}_k^{'} \mathbf{G}_k^{'} \mathbf{G}_k \mathbf{z}_k = \mathbf{z}_k^{'} \mathbf{L}_k \mathbf{z}_k = 1$, is required to be taken into account. Thus, the minimisation in (18) results in a constrained regression problem where Lagrange multipliers can be used to obtain optimal quantifications,

$$\mathbf{z}_k = \frac{\mathbf{L}_k^{-1} \mathbf{G}_k^{'} \mathbf{y}_k}{\sqrt{\mathbf{y}_k^{'} \mathbf{G}_k^{'} \mathbf{L}_k^{-1} \mathbf{G}_k^{'} \mathbf{y}_k}}. \tag{19}$$

It needs to be noted that extra care should be given to categorical variables where a natural order is required to be maintained in the categorical levels (i.e. ordinal variables). In cases

where the optimal quantifications are not naturally ordered, an ordering constraint can be imposed by introducing ties[4]. For example, suppose that the natural order of an ordinal variable is $a < b < c$ with resulting optimal quantifications $z_b < z_a < z_c$. Since $z_b < z_a$, a tie is introduced so that $z_a = z_b < z_c$.

Furthermore, Gower et al. (2011) states that solutions for categorical PCA are not nested. For example, the optimal solution for $r = 4$ does not include the solution at $r = 3$ and hence, at each dimension an optimal solution must be computed. Moreover, once convergence is obtained for an optimal solution, the resulting matrix becomes a numerical data matrix denoted as $\mathbf{H}^*_{\mathbf{cat}}$ from which a PCA biplot can be constructed.

Thus, it is clear that catPCA expands on the more traditional biplot construction techniques by allowing the use of categorical variables. It does, however, not optimally discriminate between classes as in the case of CVA. Hence, the next section will introduce a second method of incorporating categorical variables on axes in a biplot setting. This method will attempt to introduce some aspects of CVA which will be shown to improve classification performance in biplots.

3. Methodology

3.1. Introduction

In this section a new method to construct CVA biplots in instances where categorical variables are present will be proposed. The goal of this technique is, through the use of non-linear PCA (Michailidis and De Leeuw, 1998), to convert the original data matrix $\mathbf{X}$ into a combination of indicator matrices $\mathbf{G}$, and numeric vectors $\mathbf{z}$, such that $\mathbf{X} \equiv \mathbf{H} = \mathbf{Gz}$. The numeric vectors $\mathbf{z}$ are solved iteratively so that the resulting matrix, denoted $\mathbf{H}_r$, where $r =$rank$(\mathbf{H}_r)$, is the best rth dimensional PCA representation of the $\mathbf{H}$ matrix. Thereafter, the resulting $\mathbf{H}_r$ matrix is used as an input matrix to construct a CVA biplot. This new method will be referred to in this paper as CVA(H_r). The technique of CVA(H_r) is discussed in greater detail below. Thereafter, the context of the Iris data set will be presented. This section will conclude with a discussion on which biplots will be constructed on the Iris data set.

3.2. CVA(H_r)

Assuming, as before, that the original data matrix $\mathbf{X} : n \times p$ can be written in the form of $\mathbf{X} \equiv \mathbf{H} = \mathbf{Gz}$, where

$$\mathbf{H} = [\mathbf{G_1 z_1}, \mathbf{G_2 z_2}, \ldots, \mathbf{G_p z_p}]\,. \tag{20}$$

The optimal z-quantifications are then obtained in the same iterative manner as in the case of catPCA through solving the objective function given in (16) and considering the solution

[4]A technique which makes use of monotone regression to ensure a natural order is maintained

to the constraint regression problem in (19). However, in CVA(H_r), $\mathbf{Y_{cat}} = \mathbf{UDV_{[r]}}$ is the SVD of the centred matrix $\mathbf{H}$ with rank($\mathbf{Y_{cat}}$) $= r \leq p$. In contrast to catPCA, r is not required to be set equal to two since an additional dimension reduction procedure is performed at a later stage. Consequently, a pseudo-numeric representation is obtained denoted $\mathbf{H}_r$. However, in the case of CVA(H_r), rank($\mathbf{Y_{cat}}$) $= r = 2, 3, \ldots, p$, will also be considered. It is worth noting that if $\mathbf{Y_{cat}}$ is full-rank, i.e. $r = p$, the z-quantifications are simply rotated to the principle axes. The above differs compared to catPCA where only the case of $r = 2$ is considered. Furthermore, the final z-quantifications might differ when treating categorical variables as ordinal compared to nominal. The reason for this is due to the possible introduction of ties in ordinal variables to preserve the natural ordering of variables. As a result, the optimal matrix $\mathbf{H}_r$ will differ if a particular categorical variable is considered nominal versus ordinal and ties are present in the ordinal case.

Upon achieving an optimal pseudo-numeric form of $\mathbf{X}$, a CVA biplot can be constructed by transforming the resulting $\mathbf{H}_r$ matrix into the canonical space through $\mathbf{S} = \mathbf{H}_r\mathbf{L}$ with $\mathbf{L}'\mathbf{L} = \mathbf{W^{-1}}$. Thereafter, PCA can be used to approximate the canonical means $\widehat{\bar{\mathbf{H}}}_r\mathbf{L}$ with $\widehat{\bar{\mathbf{S}}} = \widehat{\bar{\mathbf{H}}}_r\mathbf{L} = \mathbf{UDV_{[2]}}$. It is worth noting that the SVD performed on $\widehat{\bar{\mathbf{S}}}$ provides the opportunity to consider higher ranks of $\mathbf{Y_{cat}}$, since the rank of the canonical means is ultimately reduced to the case where rank($\mathbf{UDV_{[2]}}$) $= s = 2$ which allows points to be plotted in the canonical space. Similarly to the case in the PCA biplot, the class means can be approximated by $\bar{\mathbf{P}} = \bar{\mathbf{S}}\mathbf{V}'_{\mathbf{[2]}} = \bar{\mathbf{H}}_r\mathbf{L}\mathbf{V}'_{\mathbf{[2]}} = \bar{\mathbf{H}}_r\mathbf{M}'_{\mathbf{[2]}}$ where $\mathbf{M}'_{\mathbf{[2]}} : p \times 2$ is the normalised eigenvector of $\mathbf{LV}'_{\mathbf{[2]}}$ and is used in constructing the axes of the respective variables.

3.3. Iris data set

The Iris data set is one of the most widely used data sets for showcasing new supervised learning modelling techniques. In particular, the data set is ubiquitous in showcasing the application of LDA. Introduced by Fisher (1936), the data set was initially used to quantify the morphologic variation of three species of Iris flowers. The data set contains 50 samples from each of the three species (Iris Setosa, Iris Virginica, and Iris Versicolor). In addition, each sample consists of four recorded features of the flower namely, the length and width of the sepals and petals measured in centimetres. A summary of descriptive statistics for the Iris data set can be found in table (A.1).

In order to introduce categorical variables in the data set, the last two variables namely, `petal length` and `petal width`, were transformed into categorical variables through binning so that each category level is equal in width[5]. In this instance the categories were split into five category levels namely, `Very Small`, `Small`, `Average`, `Large` and `Very Large`. The intervals of the binning procedure can be found in table 1.

The constructed CVA(H_r) biplots using the Iris data set where `petal length` and `petal width` are transformed into categorical variables (both nominal and ordinal cases) are given

[5]Other groupings of categorical variables were also performed with similar results.

Table 1: The label associated with the various intervals after binning is performed to transform `petal length` and `petal width` into categorical variables with five levels with each level being equal in length.

	Petal Length	Petal Width
Very Small	[0.99,2.18]	[0.09,0.58]
Small	(2.18,3.36]	(0.58,1.06]
Average	(3.36,4.54]	(1.06,1.54]
Large	(4.54,5.72]	(1.54,2.02]
Very Large	(5.72,6.91]	(2.02,2.50]

in the following section. Thereafter, the classification accuracy of the CVA(H_r) biplots are compared to that of the catPCA biplots.

4. Results

In this section, the catPCA and CVA(H_r) biplots, where rank($\mathbf{H}$) $= r = 2, 3, 4$, based on the Iris data set will be presented. This section will be split into two separate subsections. In the first subsection the case where `petal width` and `petal length` are treated as nominal variables will be considered. This will be followed by the case where `petal length` and `petal width` are treated as ordinal variables. Accuracy metrics generated from confusion matrices will be provided and discussed in detail as well. The raw confusion matrices can be found in Appendix D.

4.1. Nominal case

In the nominal case no natural order of categorical variables is required to be maintained. As a natural consequence, no ties will be introduced which allows for easier interpretation of the biplots. To further improve the interpretation of the biplot, Blasius, Eilers and Gower (2009) proposes that the axes of the categorical variables are segmented according to their various category levels, with each level assuming a different colour. The catPCA biplot along with the CVA(H_2) biplot is shown in figures 4 and 5 respectively. The CVA(H_3) and CVA(H_4) biplots are given in the appendix under figure B.1. In addition, the accuracy metrics of the various biplots can be found in table 2.

From observing the biplots, it is clear that the CVA(H_r) biplot is superior in terms of class separation. The above is confirmed when consulting the total accuracy column in table 2 where the CVA(H_2) biplot boasts a 15% increase accuracy in both the Versicolor and Virginica species classification. Even though the accuracy in the case of the CVA(H_3) and CVA(H_4) biplots is greater than that of the catPCA biplot, the accuracy of these biplots diminishes as the rank of $\mathbf{H}$ increases. The above could suggest that the most accurate

biplot representation is achieved when the rank of $\mathbf{H}$ and the dimension of the biplot are equal

It is further interesting to note that the variables that discriminate best between the classes are consistent across the two biplot construction techniques. In both biplots, `petal width` and `petal length` greatly discriminate between classes in contrast to `sepal length` which moderately discriminates between classes and `sepal width`, which offers little to no indication that the variable can separate classes.

Table 2: Classification and misclassification type errors for the catPCA and CVA(H_r) biplots where the last two variables were treated as nominal. The full data set was used in creating the classification regions. The measures were calculated as follows: Total accuracy: $\frac{TP+TN}{N}$; Positive predicted value: $\frac{1}{J}\sum_{i=1}^{J}\frac{TP_i}{PP_i}$; Negative predicted value: $\frac{1}{J}\sum_{i=1}^{J}\frac{NP_i}{PN_i}$;Sensitivity: $\frac{1}{J}\sum_{i=1}^{J}\frac{TP_i}{TP_i}$; False negative rate = $\frac{1}{J}\sum_{i=1}^{J}\frac{FN_i}{TP_i}$; False positive rate = $\frac{1}{J}\sum_{i=1}^{J}\frac{FP_i}{TN_i}$; Specificity = $\frac{1}{J}\sum_{i=1}^{J}\frac{TN_i}{TN_i}$; RSS = $\sum_{k=1}^{p}(\mathbf{G}_k\mathbf{z}_k-\mathbf{y}_k)^2$ with J equal to the number of classes, TP equal to number of true positive cases, TN equal to number of true negative cases, FN equal to the number of false negative cases, FP equal to the number of false positive cases and N equal to the total sample size.

Biplot method		Total accuracy (+)	Pos. pred. value (+)	Neg. pred. value (+)	Sensitivity (+)	False neg. rate (-)	False pos. rate (-)	Specificity (+)	RSS
catPCA	Se	0.990	1.000	0.990	0.980	0.020	0.000	1.000	0.343
	Ve	0.780	0.692	0.857	0.720	0.280	0.160	0.840	
	Vi	0.780	0.714	0.852	0.700	0.300	0.140	0.860	
CVA(H_2)	Se	0.990	1.000	0.990	0.980	0.020	0.000	1.000	0.343
	Ve	0.935	0.902	0.960	0.920	0.080	0.050	0.950	
	Vi	0.940	0.920	0.960	0.920	0.080	0.040	0.960	
CVA(H_3)	Se	1.000	1.000	1.000	1.000	0.000	0.000	1.000	0.011
	Ve	0.855	0.973	0.876	0.720	0.280	0.010	0.990	
	Vi	0.920	0.778	0.989	0.980	0.020	0.140	0.860	
CVA(H_4)	Se	1.000	1.000	1.000	1.000	0.000	0.000	1.000	0.000
	Ve	0.800	0.778	0.857	0.700	0.300	0.100	0.900	
	Vi	0.825	0.727	0.895	0.800	0.200	0.150	0.850	

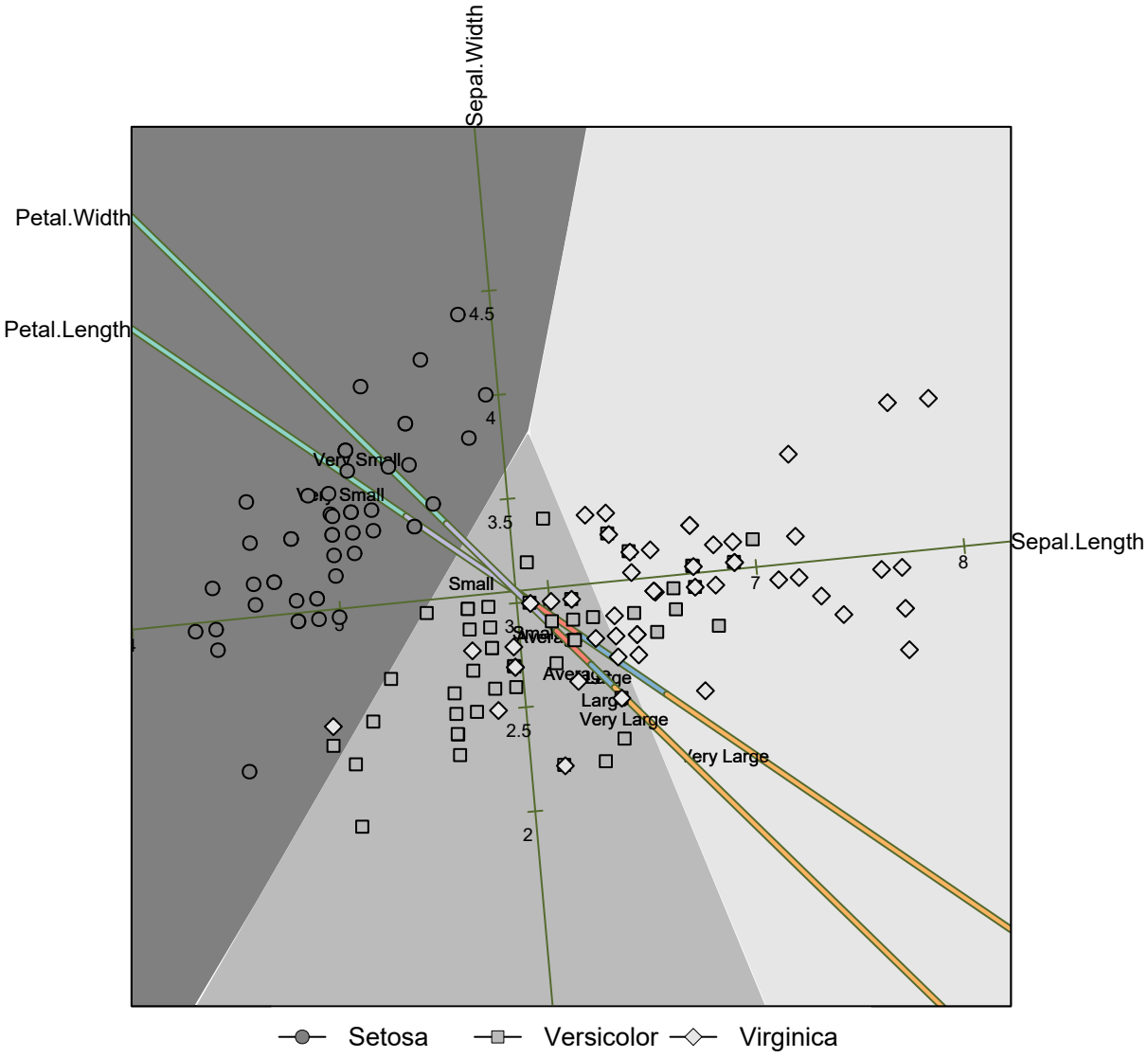


Figure 4: catPCA biplot compiled using the `catPCA` function in `R` with the categorical variables treated as nominal. The class region areas were created using an LDA model which was trained on the resulting points of the biplot.

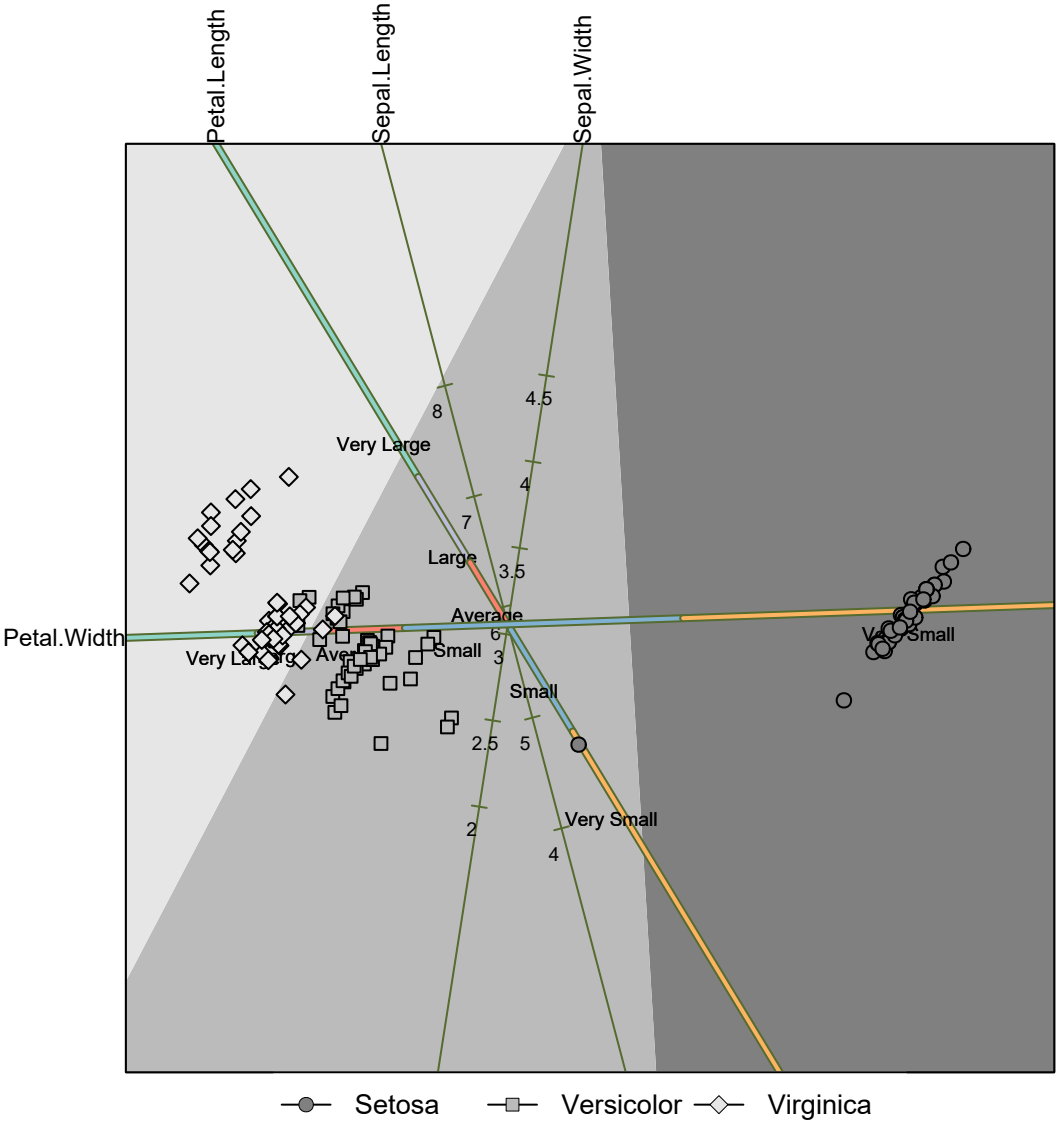


Figure 5: CVA(H_r) biplot compiled with rank($\mathbf{H}$) = 2 with the categorical variables treated as nominal.The class region areas were created using an LDA model which was trained on the resulting points of the biplot.

4.2. Ordinal case

In the ordinal case, the natural order of categorical variables must be maintained. Thus, in the case where the final z-quantifications do not preserve this natural order, ties will be introduced which may hinder the overall interpretation of the biplots. To further improve the interpretation of the biplot Blasius et al. (2009) proposes that axes of the ordinal categorical variables are segmented according to different thickness levels across the various category levels. The segmentation is performed so that the lowest ordered level is given the thinnest segment while the highest ordered level is given the thickest segment. The above allows for an intuitive understanding when referring to the ordinal category axes. The catPCA biplot along with the CVA(H_2) biplot is shown in figures 6 and 7 respectively. The CVA(H_3) and CVA(H_4) biplots are given in the appendix under figure B.2. In addition, the accuracy metrics of the various biplots can be found in table 4.

Again, as in the nominal case, the CVA(H_r) biplot technique is superior in terms of class separation and accuracy. The overall classification accuracy difference between the CVA(H_2) biplot and catPCA biplot is further increased to 15.5% and 16% for the Versicolour and Virginica classes respectively. The pattern of decreasing accuracy persists, however, when the rank of $\mathbf{H}$ increases. Also, in the ordinal case, two additional concerns are apparent, namely the introduction of ties and the collapsing of the categorical axes as the rank of $\mathbf{H}$ increases. The variables that discriminate between classes, however, remain consistent between the catPCA and CVA(H_r) biplots as with the nominal case.

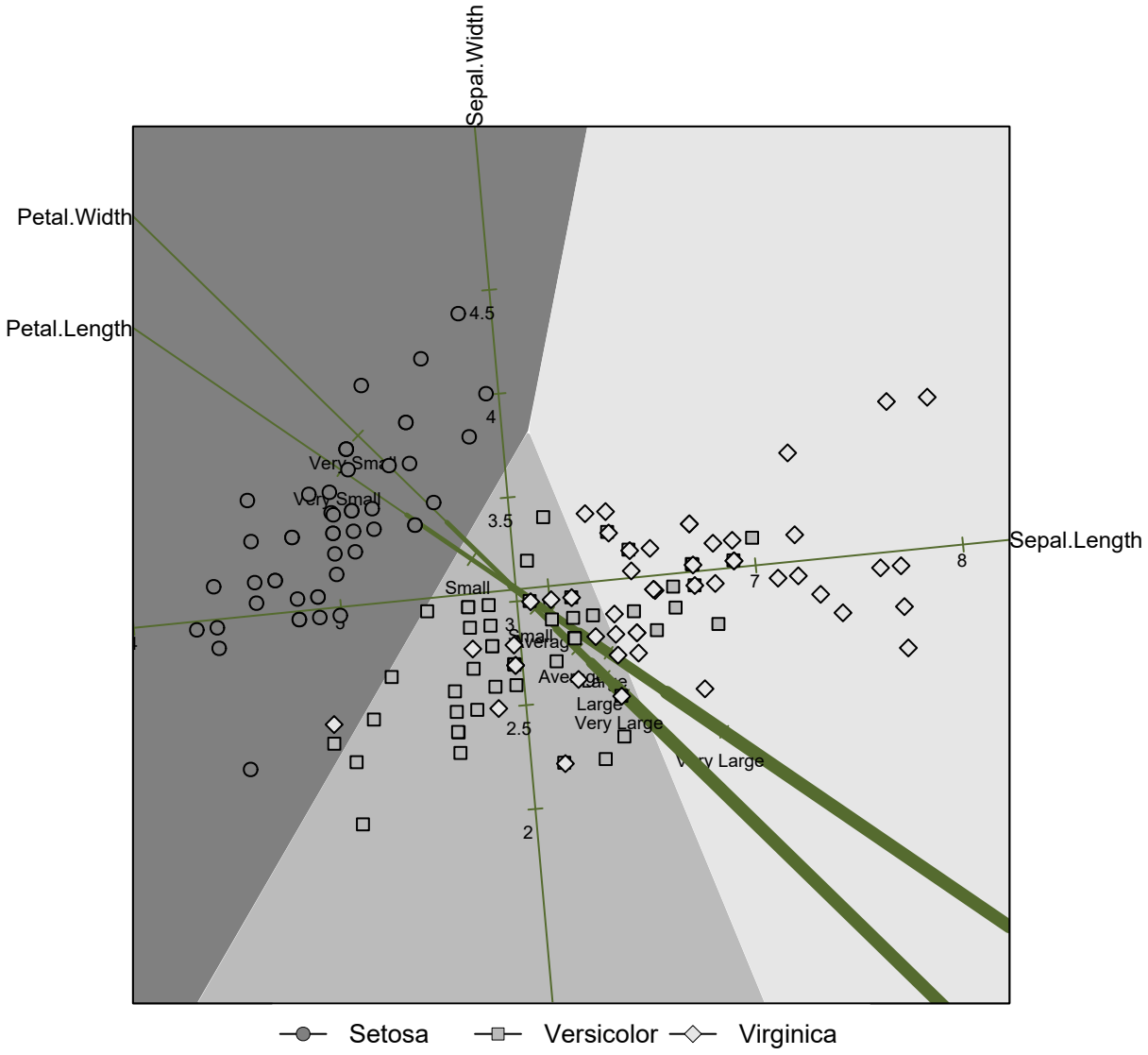


Figure 6: catPCA biplot compiled using the `catPCA` function in `R` with the categorical variables treated as ordinal. The class region areas were created using an LDA model which was trained on the resulting points of the biplot.

To observe the the reason for the introduction of additional ties, the final z-quantifications for the CVA(H_r) biplots can be observed in Appendix C. In comparing the z-quantifications obtained in the third row (rank($\mathbf{H}$) = 4) in the two figures, it is clear that the z-quantifications in the ordinal case tends to converge at the higher-ordered levels compared to the nominal case, where the z-quantifications follows a more quadratic curve. Hence, to preserve the natural order of the ordinal variables, ties are used which results in the convergence of the z-quantifications. Even though ties were only introduced in the CVA(H_4) biplot, the nominal z-quantifications exhibits a quadratic curve which becomes more pronounced at higher ranks of $\mathbf{H}$, thus in turn, increases the possibility of ties.

Moreover, when comparing the differences between the z-quantifications at each rank between `petal length` and `petal width`, it is clear that the difference between these two z-quantifications decreases dramatically as the rank of $\mathbf{H}$ increases. The exact z-quantifications in the case of the CVA(H_4) biplot is given in table 3. Consequently, this results in the orientation of the axis being almost identical since the matrix determining the direction of the axis namely, $\mathbf{L}$, is solved by considering an input of $\mathbf{H}_r$ with two almost identical columns.

Table 3: The differences of the final z-quantifications in the CVA(H_4) biplot where the `petal width` and `petal length` are treated as ordinal variables

	Very Small	Small	Average	Large	Very Large
Petal Length (1)	-0.11518	0.019235	0.05596	0.06030	0.06030
Petal Width (2)	-0.11660	0.021311	0.05960	0.05960	0.05960
Difference: (2) - (1)	-0.00141	0.00207	0.00363	-0.00070	-0.00070

4.3. Comparison between catPCA and CVA(H_r)

In comparing the two biplot construction techniques it is clear that in terms of accuracy, that the CVA(H_r) biplots are advantageous. However, it should be noted that the choice of what rank of $\mathbf{H}$ to use should be empirically tested by considering all ranks to determine which CVA(H_r) provides a biplot which produces: 1) the least ties, 2) is easy to interpret, and 3) is the most accurate in terms of class classification. In light of this, an additional advantage of CVA(H_r) biplots can be its inherent flexibility since multiple biplots can be constructed using the same data set.

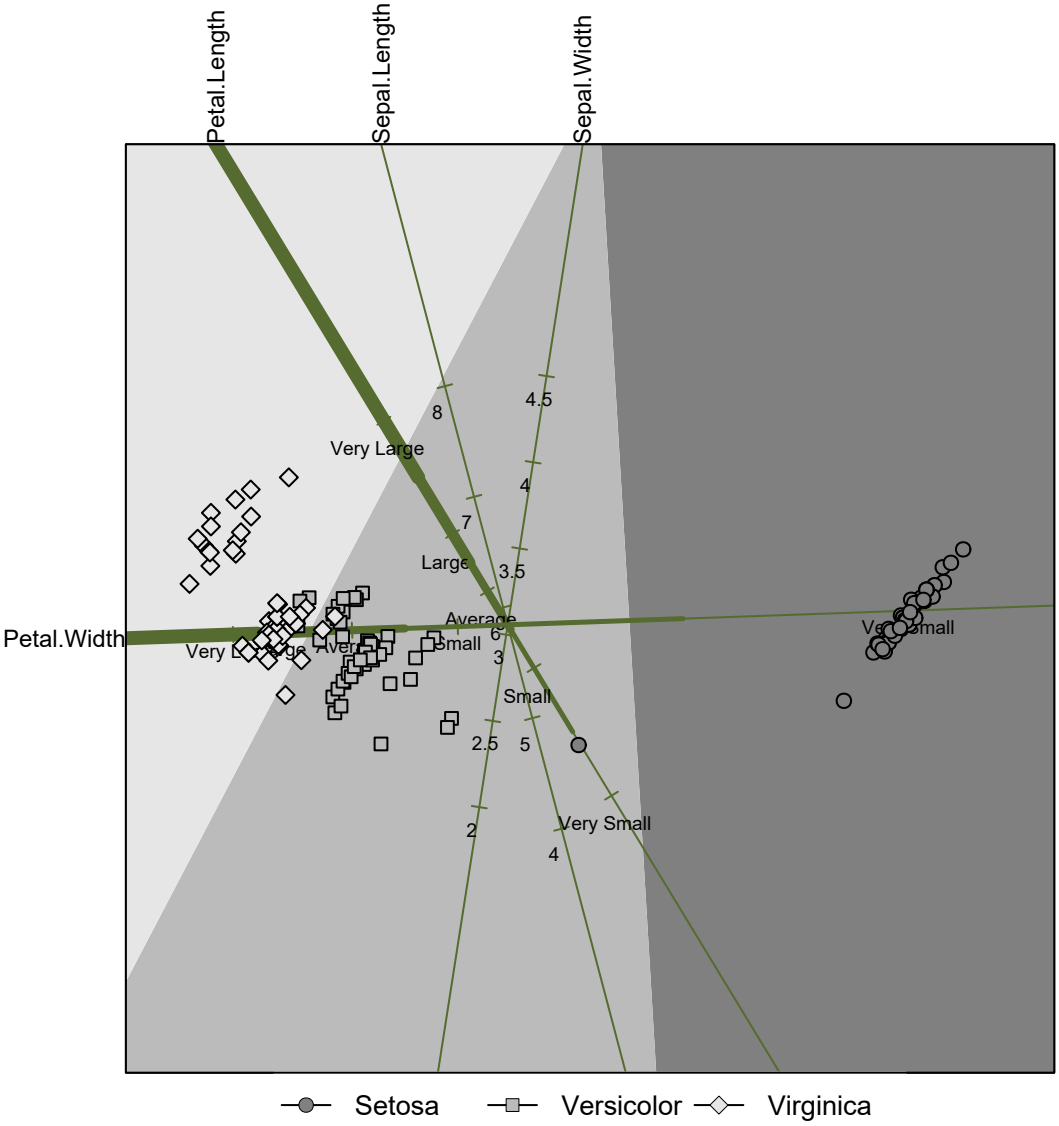


Figure 7: CVA(H_2) biplot compiled with the categorical variables treated as ordinal.The class region areas were created using an LDA model which was trained on the resulting points of the biplot.

Table 4: Classification and misclassification type errors for the catPCA and CVA(H_r) biplots where the last two variables were treated as ordinal. The full data set was used in creating the classification regions. The measures were calculated as follows: Total accuracy: $\frac{TP+TN}{N}$; Positive predicted value: $\frac{1}{J}\sum_{i=1}^{J}\frac{TP_i}{PP_i}$; Negative predicted value: $\frac{1}{J}\sum_{i=1}^{J}\frac{NP_i}{PN_i}$;Sensitivity: $\frac{1}{J}\sum_{i=1}^{J}\frac{TP_i}{TP_i}$; False negative rate = $\frac{1}{J}\sum_{i=1}^{J}\frac{FN_i}{TP_i}$; False positive rate = $\frac{1}{J}\sum_{i=1}^{J}\frac{FP_i}{TN_i}$; Specificity = $\frac{1}{J}\sum_{i=1}^{J}\frac{TN_i}{TN_i}$; RSS = $\sum_{k=1}^{p}\left(\mathbf{G}_k\mathbf{z}_k-\mathbf{y}_k\right)^2$ with J equal to the number of classes.

Biplot method		Total accuracy (+)	Pos. pred. value (+)	Neg. pred. value (+)	Sensitivity (+)	False neg. rate (-)	False pos. rate (-)	Specificity (+)	RSS
catPCA	Se	0.990	1.000	0.990	0.980	0.020	0.000	1.000	0.343
	Ve	0.780	0.692	0.857	0.720	0.280	0.160	0.840	
	Vi	0.780	0.714	0.852	0.700	0.300	0.140	0.860	
CVA(H_2)	Se	0.990	1.000	0.990	0.980	0.020	0.000	1.000	0.343
	Ve	0.935	0.902	0.960	0.920	0.080	0.050	0.950	
	Vi	0.940	0.920	0.960	0.920	0.080	0.040	0.960	
CVA(H_3)	Se	1.000	1.000	1.000	1.000	0.000	0.000	1.000	0.011
	Ve	0.855	0.973	0.876	0.720	0.280	0.010	0.990	
	Vi	0.920	0.778	0.989	0.980	0.020	0.140	0.860	
CVA(H_4)	Se	1.000	1.000	1.000	1.000	0.000	0.000	1.000	0.000
	Ve	0.810	0.783	0.865	0.720	0.280	0.100	0.900	
	Vi	0.830	0.741	0.896	0.800	0.200	0.140	0.860	

5. Conclusion

This paper expands on multivariate visualisation techniques to incorporate categorical variables in a biplot setting. A new method is proposed using both non-linear PCA and CVA called CVA(H_r). These biplots provide an intuitive visualisation with more depth and are easier to interpret than black-box methods. The method above was subsequently tested using the well-known IRIS data set where it greatly improves classification accuracy and introduces greater flexibility in contrast to existing methods. An area which is identified as possible areas for further research is to attempt to incorporate CVA characteristics in the constraint regression problem which plays the role of determining optimal z-quantifications. Furthermore, it will be of great interest to observe the performance of these biplots in other real-world applications.

References

Aldrich, C., Gardner, S., Le Roux, N., 2004. Monitoring of metallurgical process plants by using biplots. AIChE Journal 50, 2167–2186. doi:10.1002/aic.10170.

Alkan, B.B., Atakan, C., Akdi, Y., 2015. Visual analysis using biplot techniques of rainfall changes over turkey. Mapan 30, 25–30. doi:10.1007/s12647-014-0119-8.

Blasius, J., Eilers, P., Gower, J., 2009. Better biplots. Computational Statistics & Data Analysis 53, 3145–3158. doi:10.1016/j.csda.2008.06.013.

Campbell, N.A., Atchley, W.R., 1981. The Geometry of Canonical Variate Analysis. Systematic Biology 30, 268–280. doi:10.1093/sysbio/30.3.268.

Fisher, R.A., 1936. The use of multiple measurements in taxonomic problems. Annals of eugenics 7, 179–188. doi:10.1111/j.1469-1809.1936.tb02137.x.

Gardner, S., Le Roux, N., 2005. An identification biplot for detecting forgery. .

Gower, J.C., Hand, D.J., 1996. Biplots. volume 54. CRC Press.

Gower, J.C., Lubbe, S., Le Roux, N.J., 2011. Understanding biplots. John Wiley & Sons.

Greenacre, M.J., 2010. Biplots in practice. Fundacion BBVA. URL: http://www.fbbva.es.

Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. Journal of educational psychology 24, 417. doi:10.1037/h0071325.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An introduction to statistical learning. volume 112. Springer.

Jolliffe, I.T., Cadima, J., 2016. Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 374, 20150202. doi:10.1098/rsta.2015.0202.

Michailidis, G., De Leeuw, J., 1998. The gifi system of descriptive multivariate analysis. Statistical Science , 307–336doi:10.1214/ss/1028905828.

Rao, C.R., 1948. The utilization of multiple measurements in problems of biological classification. Journal of the Royal Statistical Society. Series B (Methodological) 10, 159–203.

Rao, C.R., 1952. Advanced statistical methods in biometric research. New York , 351–382doi:10.1002/ajpa.1330120224.

Van der Merwe, C.J., 2020. Classifying yield spread movements in sparse data through triplots. Ph.D. thesis. Ghent University & Stellenbosch University. URL: http://hdl.handle.net/1854/LU-8643411.

Varas, M., Vicente-Tavera, S., Molina, E., Vicente-Villardón, J., 2005. Role of canonical biplot method in the study of building stones: an example from spanish monumental heritage. Environmetrics: The official journal of the International Environmetrics Society 16, 405–419. doi:10.1002/env.722.

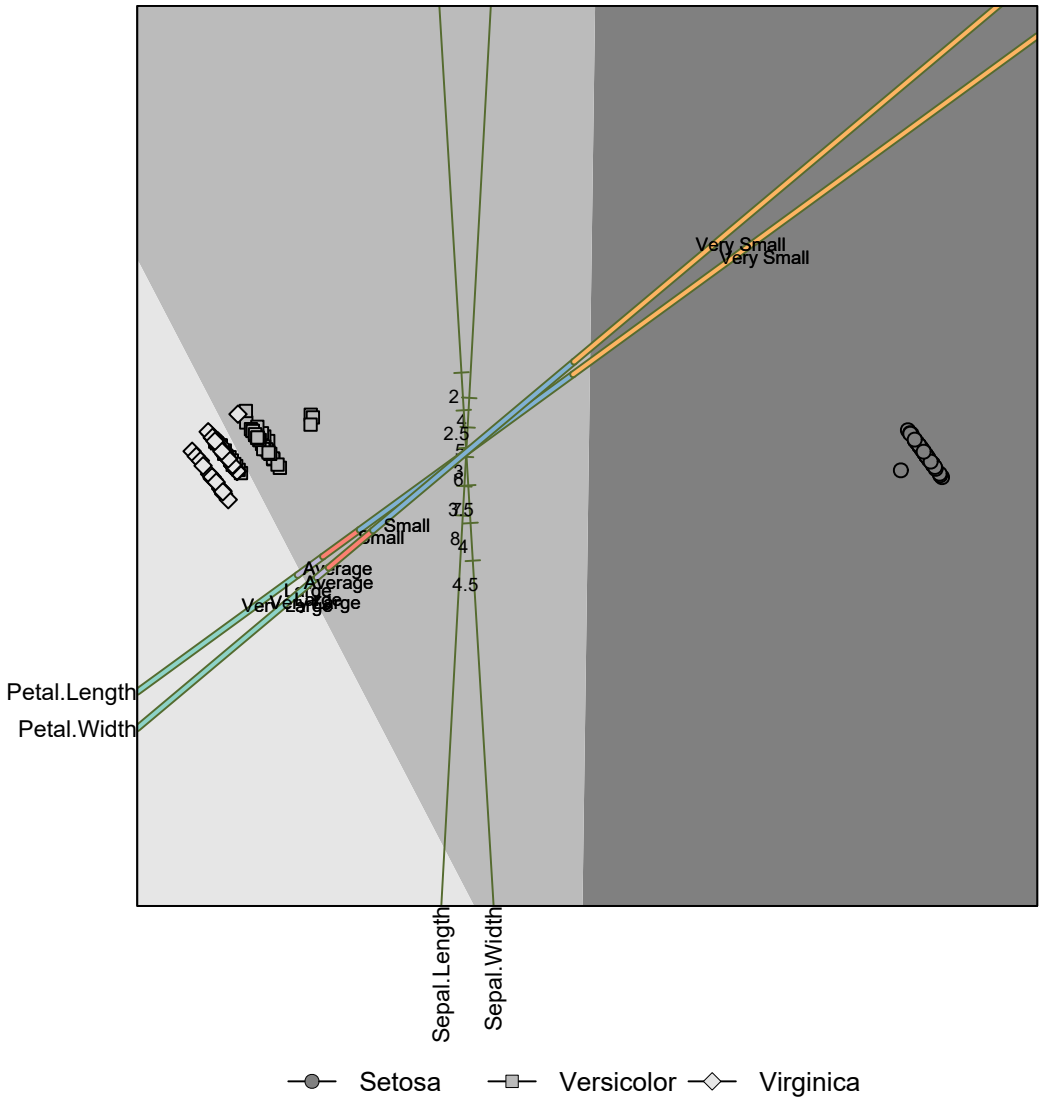
Walters, I., Le Roux, N., 2008. Monitoring gender remuneration inequalities in academia using biplots. ORiON 24, 49–73. doi:10.5784/24-1-59.

Appendix A Descriptive statistics of the Iris data set

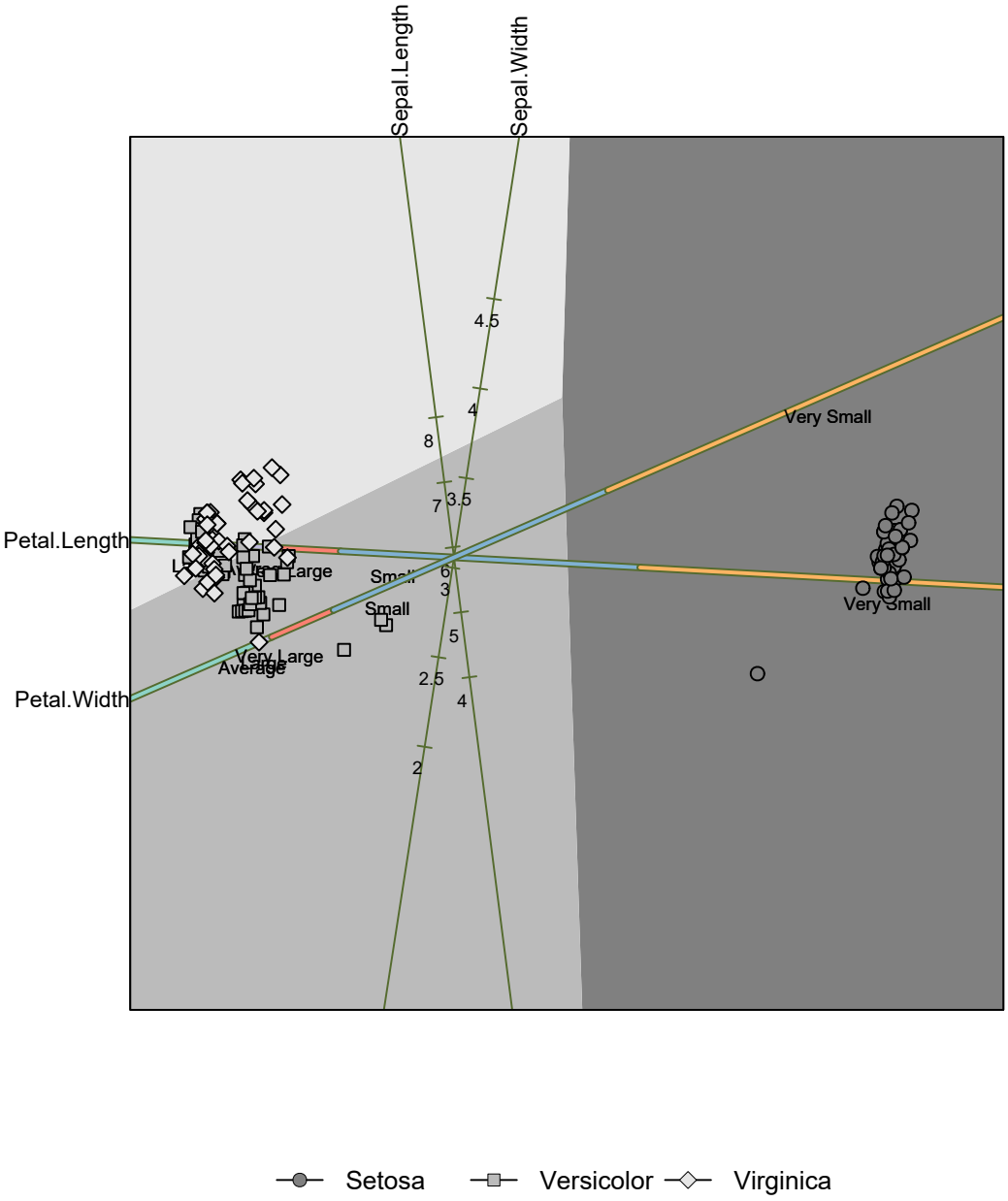
Table A.1: Descriptive statistics of the four morphologic features split into each of the three flower species for the Iris data set used to construct biplots.

Variable	Class	Mean	Std. Dev.	Med.	Min	Max	25th perc.	75th perc.	Skew-ness	Kurt-osis
Sepal Length	Setosa	5.006	0.352	5.000	4.300	5.800	4.800	5.200	0.113	-0.451
	Versicolor	5.936	0.516	5.900	4.900	7.000	5.600	6.300	0.099	-0.694
	Virginica	6.588	0.636	6.500	4.900	7.900	6.225	6.900	0.111	-0.203
Sepal Width	Setosa	3.428	0.379	3.400	2.300	4.400	3.200	3.675	0.039	0.596
	Versicolor	2.770	0.314	2.800	2.000	3.400	2.525	3.000	-0.341	-0.549
	Virginica	2.974	0.322	3.000	2.200	3.800	2.800	3.175	0.344	0.380
Petal Length	Setosa	1.462	0.174	1.500	1.000	1.900	1.400	1.575	0.100	0.654
	Versicolor	4.260	0.470	4.350	3.000	5.100	4.000	4.600	-0.571	-0.190
	Virginica	5.552	0.552	5.550	4.500	6.900	5.100	5.875	0.517	-0.365
Petal Width	Setosa	0.246	0.105	0.200	0.100	0.600	0.200	0.300	1.180	1.259
	Versicolor	1.326	0.198	1.300	1.000	1.800	1.200	1.500	-0.029	-0.587
	Virginica	2.026	0.275	2.000	1.400	2.500	1.800	2.300	-0.122	-0.754

Appendix B CVA(H_r) Biplots

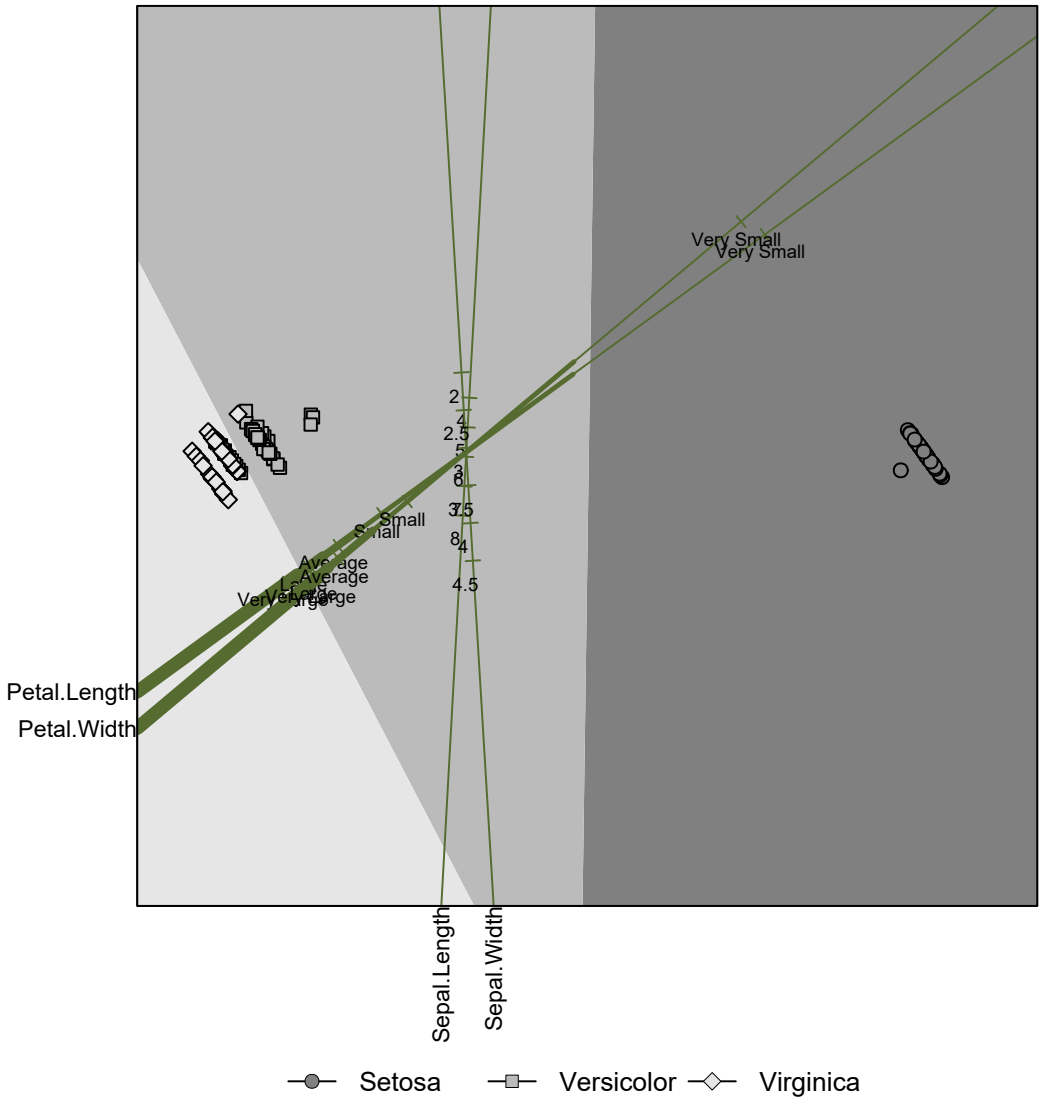


(a) A CVA(H_3) biplot. The class region areas were created using an LDA model which was trained on the resulting points of the biplot.



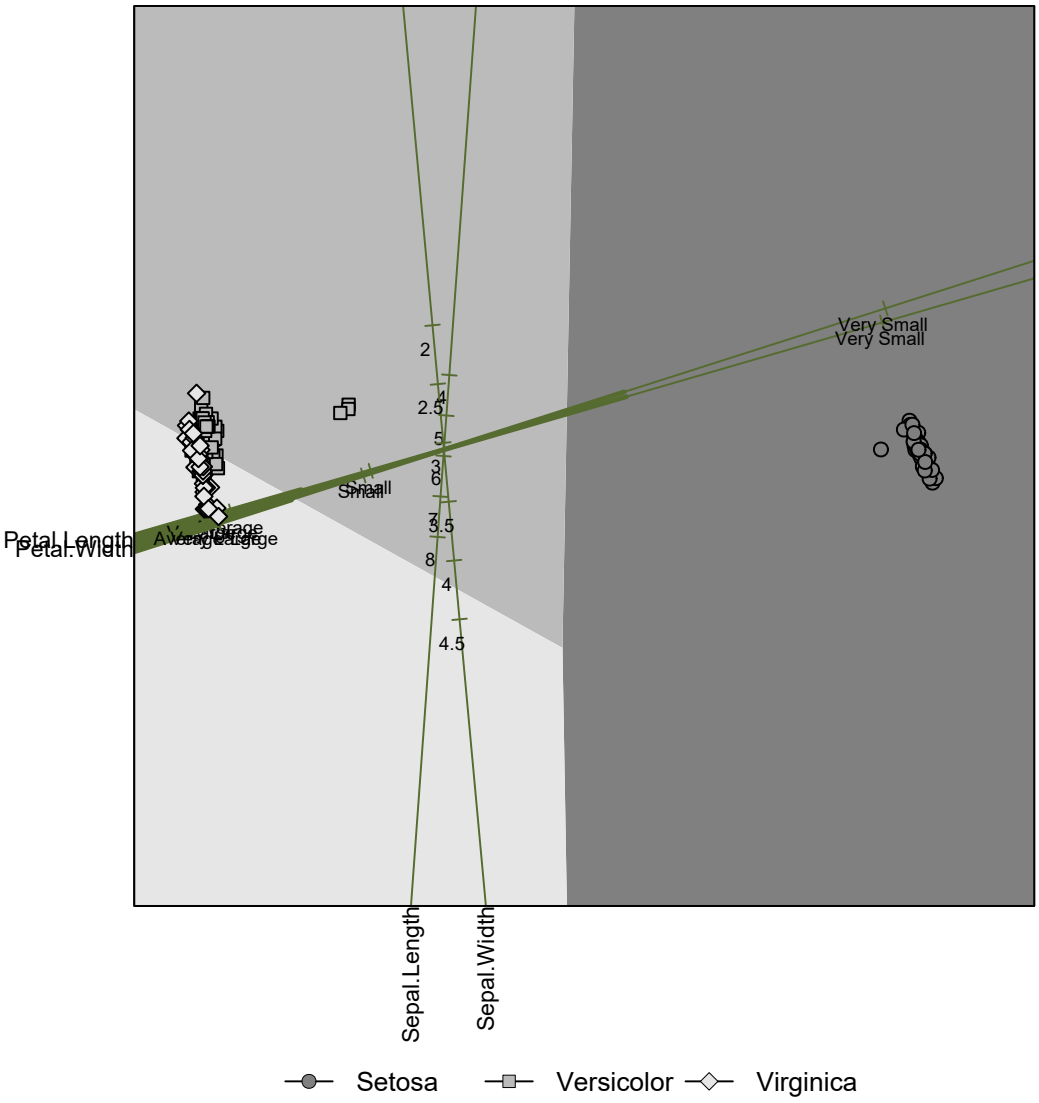
(b) A CVA(H_4) biplot. The class region areas were created using an LDA model which was trained on the resulting points of the biplot.

Figure B.1: CVA(H_r) biplots where `petal length` and `petal width` are treated as nominal variables (cont.)



(a) A CVA(H_3) biplot. The class region areas were created using an LDA model which was trained on the resulting points of the biplot.

Figure B.2: CVA(H_r) biplots where `petal length` and `petal width` are treated as ordinal variables



(b) A CVA(H_4) biplot. The class region areas were created using an LDA model which was trained on the resulting points of the biplot.

Figure B.2: CVA(H_r) biplots where `petal length` and `petal width` are treated as ordinal variables (cont.)

Appendix C Final z quantifications

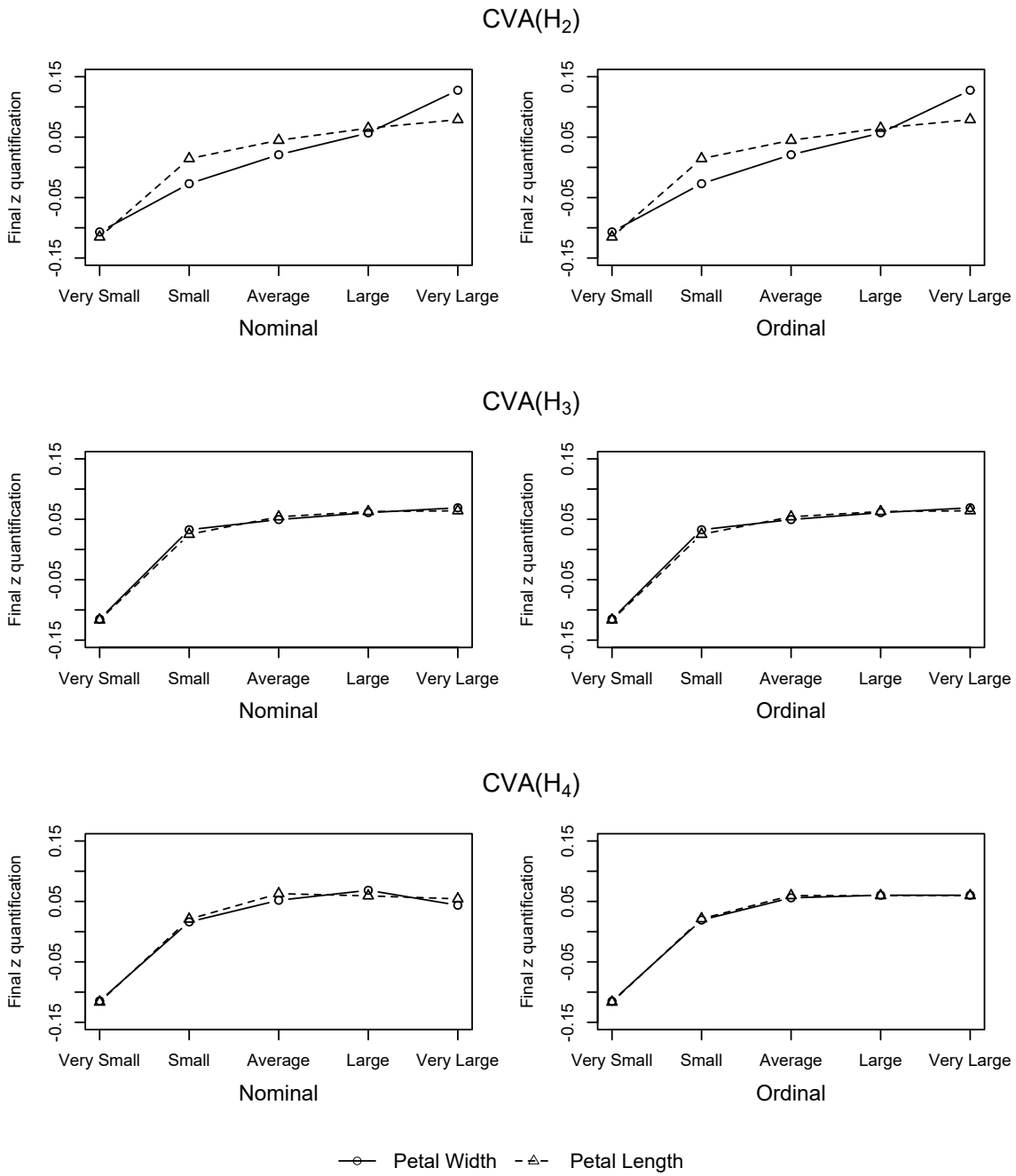


Figure C.3: Final **z** quantifications for the CVA(H_r) biplots. Each row corresponds to the case where the rank of $\mathbf{H} = r = 2, 3, 4$.

Appendix D Confusion matrices

These confusion matrices were generated using an LDA model that was trained on the points of the respective biplots. The full set of points were used for training and prediction.

Table D.2: Confusion matrices in the case where `petal width` and `petal length` are treated as categorical variables

NOMINAL CASE

catPCA	REFERENCE			**CVA(H_2)**	REFERENCE		
PREDICTION	Setosa	Versicolor	Virginica	PREDICTION	Setosa	Versicolor	Virginica
Setosa	49	0	0	Setosa	49	0	0
Versicolor	1	36	15	Versicolor	1	46	4
Virginica	0	14	35	Virginica	0	4	46
CVA(H_3)	REFERENCE			**CVA(H_4)**	REFERENCE		
PREDICTION	Setosa	Versicolor	Virginica	PREDICTION	Setosa	Versicolor	Virginica
Setosa	50	0	0	Setosa	50	0	0
Versicolor	1	36	1	Versicolor	0	35	10
Virginica	0	14	49	Virginica	0	15	40

ORDINAL CASE

catPCA	REFERENCE			**CVA(H_2)**	REFERENCE		
PREDICTION	Setosa	Versicolor	Virginica	PREDICTION	Setosa	Versicolor	Virginica
Setosa	49	0	0	Setosa	49	0	0
Versicolor	1	36	15	Versicolor	1	46	4
Virginica	0	14	35	Virginica	0	4	46
CVA(H_3)	REFERENCE			**CVA(H_4)**	REFERENCE		
PREDICTION	Setosa	Versicolor	Virginica	PREDICTION	Setosa	Versicolor	Virginica
Setosa	50	0	0	Setosa	50	0	0
Versicolor	1	36	1	Versicolor	0	36	10
Virginica	0	14	49	Virginica	0	14	40